Tail distribution of the delay in a general batch-service queueing model

Dieter Claeys, Bart Steyaert, Joris Walraevens[1], Koenraad Laevens, Herwig Bruneel

Stochastic Modelling and Analysis of Communication Systems (SMACS) Research Group, Department of Telecommunications and Information Processing (TELIN), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Tel.: +32 9 264 3411
Fax: +32 9 264 4295

Abstract

Batch servers are capable of processing batches of packets instead of individual packets. Although batch-service queueing models have been studied extensively during the past decades, the focus was mainly put on calculating performance measures related to the buffer content, whereas less attention has been paid to the packet delay. In this paper, we focus on the tail probabilities of the delay that a random packet experiences in a general batch-service queueing model. More specifically, we establish approximations for these probabilities, which are highly accurate and easy to calculate. These results, for instance, allow to accurately assess the probability that real-time packets experience an excessive delay in practical telecommunication systems.

Keywords: queueing, batch service, batch arrivals, service threshold, delay, tail probabilities
PACS: 60K25, 68M20, 90B22

1. Introduction

Whereas servers in traditional queueing systems serve one packet at a time, batch servers process batches of packets. The maximum number of

Email address: Dieter.Claeys@telin.ugent.be (Dieter Claeys)
[1]The third author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

packets in a served batch is usually finite and is called the server capacity, which we denote by c. An inherent feature of batch service is that newly arriving packets cannot join the ongoing service, even if the served batch is not completely filled. In order to reduce the wasted capacity, one often imposes a threshold, l ($1 \leq l \leq c$), for the minimum amount of packets in a served batch. This implies that the available server solely initiates service when at least l packets have accumulated in the system.

Batch-service queueing models have a wide area of applications, including transportation, production and manufacturing systems (see e.g. [7], [17]) and telecommunications (see e.g. [2]). Batch-service queueing models are for instance employed to assess the performance of burst-frame-based MAC protocols for ultra-wideband (UWB) Wireless Personal Area Networks (WPANs) ([25]). A node in such a network typically has for each combination of destination and Quality of Service (QoS) an output and a transmission buffer. Upper-layer packets with the same destination and QoS are stored in the same output buffer. When the transmission buffer is empty and at least l packets have accumulated in the output buffer, maximum c of these packets are grouped into a burst and this burst is stored in the transmission buffer (note that the transmission buffer can only store one burst simultaneously). The burst will be removed from the transmission buffer when an ACK frame from the receiver arrives. Although UWB is a high-speed technology, the time spent in the transmission buffer cannot be ignored due to the competition for the channel between the several output queues and the synchronisation (process of synchronising the receivers clock with the transmitters clock) time. The batch-service queueing model in this paper can be used to model an output and transmission buffer corresponding to a particular destination and QoS: the output buffer is the queue of the batch-service queueing model, the transmission buffer is the server and the time that a burst resides in the transmission buffer is the service time. This application example thus demonstrates that the analysis of the delay in a batch-service queueing system with general service times and a general batch forming policy is important. This theoretical analysis is subject of this paper.

On account of the wide area of applications, batch-service queueing models have been studied extensively. The emphasis was laid on the amount of packets in the system (e.g. [1], [5], [6], [8], [17], [18], [19], [21], [24], [28], [29], [30], [32], [33]). The packet delay, however, has only attracted attention in [7], [13], [14], [16], [22], [23], [26] and [27]. In none of these papers, models are studied with the combination of $l > 1$ and batch arrivals.

In [9], we have computed the probability generating function (PGF) of the packet delay in a discrete-time batch-arrival, batch-service queueing model with $l = c$ and single-slot service times. In [10], we have extended this model to geometrically distributed service times and in [12] we have considered generally distributed service times and $1 \leq l \leq c$. The established PGF's, though, suffer from the drawback that they are not suitable to extract tail probabilities. However, in several cases, this is an important performance measure. For instance, consider an output buffer that stores voice packets. Voice packets are delay-sensitive, meaning that when they arrive too late at the end user (for instance after more than 150 ms), they become useless. The quality of the upperlayer conversations is expressed in terms of the (order of magnitude of the) probability of this event (see e.g. [15]).
In view of this, we have established in [11] an approximation for the tail probabilities of the delay that a random packet experiences in the batch-arrival, batch-service queueing model with single-slot service times and $l = c$. In this paper, we extend this previous research by considering the extended model with $l \in [1, c]$ and generally distributed service times. In addition, we also obtain another approximation that allows us to more accurately assess the delay performance in the batch-service queueing model under study. The paper is organised as follows: the model is described in detail in section 2. The approximations are established in section 3, while in section 4, we demonstrate through some examples that these are highly accurate. Hence, the approximation formulas can be adopted to accurately assess the delay performance in practical batch-service queueing systems.

2. Model

In this paper, we consider a discrete-time queueing model. Packets arrive one by one and several packets can arrive in a slot. We call this batch arrivals. The number of packet arrivals during consecutive slots is generated by an independent and identically distributed (IID) process. The number of packet arrivals during slot k is denoted by A_k; A represents the number of packet arrivals during a random slot and its PGF is denoted by $A(z)$.
The number of packets in a served batch is upper-bounded by the server capacity c and lower-bounded by the threshold l ($1 \leq l \leq c$), implying that when the server becomes available and finds less than l packets, he waits to initiate service and leaves the already present packets in the queue until the beginning of the first slot whereby at least l packets have accumulated in the

system. When the system contains more than c packets at that time, the server only processes the first c packets and leaves the others in the queue (according to the first-come-first-served policy). Consecutive batch service times do not depend on the number of packets in the served batches, nor on the number of packet arrivals and they constitute an IID process. The service time of any batch is designated by T and its associated PGF by $T(z)$.

The results obtained in this paper are valid under the following assumptions:

Assumption 1. *the load* $\rho \triangleq \mathrm{E}[A]\,\mathrm{E}[T]/c < 1$*;*

Assumption 2. $R > 1$*, with* R *the radius of convergence of* $T(A(z))$*;*

Assumption 3. $\lim_{z\uparrow R} T(A(z))/z^c > 1$*;*

Assumption 4. $z^c - T(A(z))$ *is aperiodic, meaning that the highest common factor of the set of integers* $\left\{\{c\} \cup \left\{n \in \mathbb{N} : \frac{d^n}{dz^n} T(A(z))\Big|_{z=0} \neq 0\right\}\right\}$ *equals 1.*

Note that assumption 2 implies that $R_A > 1$ and $R_T > 1$ with R_A and R_T the radii of convergence of $A(z)$ and $T(z)$ respectively. Further, assumption 3 is always fulfilled if $T(A(z))$ has a finite pole R. Vice versa, if assumption 3 is not fulfilled, then R necessarily is a branch point of $T(A(z))$, and a seperate ad-hoc analysis of the packet delay tail distribution is required.

3. Deducing the approximation formulas

In order to compute the probability that the delay W (being the sojourn time in the queue) of a randomly tagged packet exceeds some large value, we split the delay into two parts. We illustrate this by means of the example depicted in Fig. 1. The tagged packet's arrival slot is denoted by J and Q_J represents the queue content (i.e. the number of packets in the queue, those in service excluded) at the beginning of slot J. Further, B (X resp.) represents the number of packet arrivals in slot J arriving before (after resp.) the tagged packet. The first part of the delay, W_1, is the time required to serve the batches with previously arrived packets. It is equal to the remaining service time of the batch being served in slot J (if any), plus the sum of $\lfloor \frac{Q_J+B}{c} \rfloor$ service times, where $\lfloor . \rfloor$ represents the floor function, i.e. $\lfloor x \rfloor = \max\{n \in \mathbb{N} \mid n \leq x\}$. Hence, in the example, $W_1 = 3$, because $T(z) = z$, $c = 10$ and

$Q_J + B = 32$. The second part, W_2, is the time until enough packets are present to fill the batch of the tagged packet with at least l packets. Mark that exactly $(Q_J + B) \bmod c$ of the previously arrived packets are served in the same batch as the tagged packet. As $l = 5$, $((Q_J + B) \bmod c) = 2$, $X = 1$, $A_{J+1} = 0$ and $A_{J+2} = 3$, W_2 takes two slots in the example. The total delay of the tagged packet then equals

$$W = \max(W_1, W_2) \ , \tag{1}$$

since the service of the tagged packet's batch can commence only if all preceding batches have been served, and the packet's batch itself contains at least l packets. Calculation of joint probabilities of W_1 and W_2 is difficult. Therefore, we propose some lower and upper bounds, that only require calculation of marginal tail probabilities of W_1 and W_2.

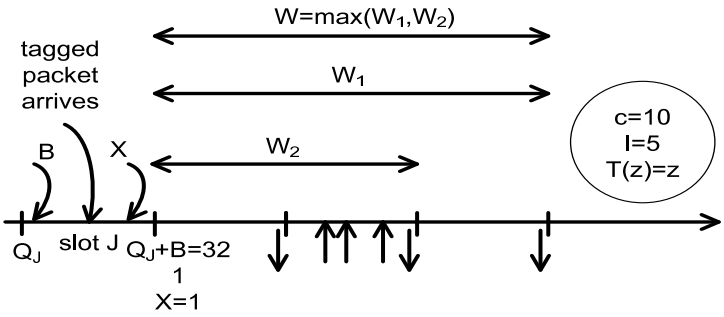


Figure 1: Illustration of W, W_1 and W_2 and introduction of some notations

On account of (1), we obtain

$$\begin{aligned} \Pr[W > w] &= \Pr[W_1 > w \vee W_2 > w] \\ &= \Pr[W_1 > w] + \Pr[W_2 > w] - \Pr[W_1 > w \wedge W_2 > w] \ . \end{aligned}$$

The following property paves the path towards establishment of a lower bound:

$$\Pr[W_1 > w \wedge W_2 > w] \leq \min(\Pr[W_1 > w], \Pr[W_2 > w]) \ . \tag{2}$$

A lower bound is obtained by assuming that the equality in (2) holds, leading to

$$\Pr[W > w] \geq \max(\Pr[W_1 > w], \Pr[W_2 > w]) \ . \tag{3}$$

An upper bound is established by assuming that $\Pr[W_1 > w \wedge W_2 > w] = 0$, leading to:

$$\Pr[W > w] \leq \Pr[W_1 > w] + \Pr[W_2 > w] \quad . \tag{4}$$

These bounds require the calculations of $\Pr[W_1 > w]$ and $\Pr[W_2 > w]$, which are discussed in the two following subsections respectively.

3.1. The calculation of $\Pr[W_1 > w]$

It was established in [12] p.27 that the PGF $W_1(z)$ of W_1 reads

$$W_1(z) = \frac{T(z)-1}{c\mathrm{E}[A]\,T(z)} \sum_{j=0}^{c-1} \frac{A\left(T(z)^{1/c}\varepsilon_j\right)-1}{\left(T(z)^{1/c}\varepsilon_j - 1\right)^2} \frac{T(z)^{1/c}\varepsilon_j}{z - A\left(T(z)^{1/c}\varepsilon_j\right)} \left\{ (z-1) \sum_{m=0}^{l-1} q_0(m) \left(T(z)^{1/c}\varepsilon_j\right)^m + \sum_{m=l}^{c-1} e(m) \left[T(z) - \left(T(z)^{1/c}\varepsilon_j\right)^m\right] \right\} \quad , \tag{5}$$

with $z^{1/c} \triangleq |z|^{1/c} e^{\imath \mathrm{Arg}(z)/c}$, whereby $\imath$ characterises the imaginary unit, $|z|$ is the absolute value of z and $\mathrm{Arg}(z)$ represents the principal value of the argument of z (i.e. it is a mapping in the interval $]-\pi, \pi]$). In addition, ε_j, $0 \leq j \leq c-1$, is the j-th complex c-th root of 1, i.e. $\varepsilon_j \triangleq e^{(\imath 2\pi j)/c}$ and $d(m)$, $0 \leq m \leq l-1$ and $e(m)$, $l \leq m \leq c-1$, are unknowns that have to be calculated by solving a set of linear equations (see [12], p.6-7).

We now compute $\Pr[W_1 > w]$ by means of the dominant-pole approximation (see e.g. [3], [4]). This technique requires that the dominant singularities (i.e. the singularities with the smallest modulus) of $W_1(z)$ are known. Unlike for the queueing system with single-slot service times and $l = c$ in [11], the dominant singularities are difficult to locate in this case. Indeed, the singularities of $W_1(z)$ might consist of zeroes of $T(z)^{1/c}\varepsilon_j - 1$ outside the closed complex unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$, zeroes of $z - A\left(T(z)^{1/c}\varepsilon_j\right)$ outside the complex unit disk, (possible) singularities of $T(z)$ and possible singularities of $A(T(z)^{1/c}\varepsilon_j)$. The following theorems play a crucial role in locating the dominant singularities.

Theorem 1. *The factor* $(T(z)^{1/c}\varepsilon_j - 1)^2$ *produces no poles* $\forall j$, $0 \leq j \leq c-1$.

Proof Suppose z_j is a zero of $T(z)^{1/c}\varepsilon_j - 1$ with multiplicity n. Then

$$\frac{\mathrm{d}^m}{\mathrm{d}z^m}(T(z)^{1/c}\varepsilon_j - 1)\Big|_{z=z_j} = 0, \forall m, 0 \leq m \leq n-1 \ ,$$

and consequently:

- $\frac{\mathrm{d}^m}{\mathrm{d}z^m}(T(z)-1)\Big|_{z=z_j} = 0, \forall m, 0 \leq m \leq n-1$, meaning that z_j is also a zero of $T(z)-1$ with multiplicity n, which appears in the numerator;
- $\frac{\mathrm{d}^m}{\mathrm{d}z^m}(A(T(z)^{1/c}\varepsilon_j)-1)\Big|_{z=z_j} = 0, \forall m, 0 \leq m \leq n-1$, meaning that z_j is also a zero of $A(T(z_j)^{1/c}\varepsilon_j)-1$ with multiplicity n, which appears in the numerator.

Summarized, although z_j is a zero of $(T(z)^{1/c}\varepsilon_j - 1)^2$ with multiplicity $2n$, it is not a pole of $W_1(z)$, since z_j is also a zero of the numerator with multiplicity $2n$. □

Lemma 1. *IF: assumptions 1-4 are satisfied,*
THEN: $z^c - T(A(z))$ has exactly one zero in the interval $]1, R[$. In addition, the zero has multiplicity one and $z^c - T(A(z))$ contains no other zeroes with a modulus larger than one and smaller than or equal to this real zero.

Proof This lemma has been proved in [31]. □

Let us denote the only zero of $z^c - T(A(z))$ in the interval $]1, R[$ by $\tilde{z}_0$. Since $\tilde{z}_0 < R \leq R_A$, the following definition makes sense:

$$z_0 \triangleq A(\tilde{z}_0) \ .$$

It holds that $z_0 \in \mathbb{R}$ and $z_0 > 1$, since $A(1) = 1$ and the PGF $A(z)$ is a real-valued and monotonically increasing function within $[1, R_A[$. In addition, $z_0 < R_T$, as $\tilde{z}_0 < R$ implies that $z_0 = A(\tilde{z}_0) < R_T$.

Theorem 2. *IF: assumptions 1-4 are satisfied,*
THEN:

1. *$T(z_0)^{1/c} < R_A$ and z_0 is a zero of $z - A(T(z)^{1/c})$;*
2. *the equations $z - A(T(z)^{1/c}\varepsilon_j)$, $0 \leq j \leq c-1$ contain no other zeroes with a modulus larger than one and smaller than or equal to z_0;*

3. *z_0 is a zero multiplicity one.*

Proof 1. On account of lemma 1, we have

$$\tilde{z}_0^c = T(A(\tilde{z}_0)) \ . \tag{6}$$

As $\tilde{z}_0$ and $T(A(\tilde{z}_0))$ are both real positive numbers, (6) can be transformed into

$$\tilde{z}_0 = T(A(\tilde{z}_0))^{1/c} \ ,$$

which is, owing to the definition of z_0, equivalent to

$$\tilde{z}_0 = T(z_0)^{1/c} \ .$$

Finally, taking into account that $T(z_0)^{1/c} = \tilde{z}_0 < R \leq R_A$ and invoking the definition of z_0, we find

$$z_0 = A(T(z_0)^{1/c}) \ .$$

In other words, $T(z_0)^{1/c} < R_A$ and z_0 is a zero of $z - A(T(z)^{1/c})$.

2. This part is a proof by contradiction. Assume that a j $(0 \leq j \leq c-1)$ exists, for which $z - A(T(z)^{1/c}\varepsilon_j)$ has a zero, z^*, with $z^* \neq z_0$ and $1 < |z^*| \leq z_0$. Owing to $|z^*| \leq z_0 < R_T$, the following definition makes sense: $\tilde{z}^* \triangleq T(z^*)^{1/c}\varepsilon_j$.
Consequently, we have that

$$|\tilde{z}^*|^c = |T(z^*)| \leq \sum_{n=1}^{\infty} \Pr[T=n] \, |z^*|^n \leq \sum_{n=1}^{\infty} \Pr[T=n] \, z_0^n = T(z_0) = \tilde{z}_0^c \ .$$

Hence, as both $|\tilde{z}^*|$ and $\tilde{z}_0$ are positive real numbers,

$$|\tilde{z}^*| \leq \tilde{z}_0 \ . \tag{7}$$

This implies that $|\tilde{z}^*| < R_A$ and taking into account that $z^* = A(T(z^*)^{1/c}\varepsilon_j)$, we find that $z^* = A(\tilde{z}^*)$. As a consequence, $|A(\tilde{z}^*)| < R_T$ and $T(A(\tilde{z}^*)) = T(z^*) = (\tilde{z}^*)^c$, meaning that $\tilde{z}^*$ is a zero of $z^c - T(A(z))$. On account of lemma 1 however, we have that $\tilde{z}_0$ is the zero with the smallest modulus larger than one of this equation and $\tilde{z}_0$ is the only zero with that modulus, so that $|\tilde{z}^*| > \tilde{z}_0$, which is a contradiction with (7).

3. The property of z_0 having multiplicity one is also a proof by contradiction. If z_0 would have a multiplicity larger than one, then (we use primes to indicate derivatives)

$$\begin{aligned} & \frac{\mathrm{d}}{\mathrm{d}z}[z - A(T(z)^{1/c})]\Big|_{z=z_0} = 0 \\ \Leftrightarrow\ & 1 - A^{'}(T(z_0)^{1/c})\frac{1}{c}T(z_0)^{1/c-1}T^{'}(z_0) = 0 \end{aligned}$$

Writing this in terms of $\tilde{z}_0$ instead of in z_0, we further transform this to

$$\begin{aligned} & c - A^{'}(\tilde{z}_0)\tilde{z}_0^{1-c}T^{'}(A(\tilde{z}_0)) = 0 \\ \Leftrightarrow\ & c\tilde{z}_0^{c-1} - T^{'}(A(\tilde{z}_0))A^{'}(\tilde{z}_0) = 0 \\ \Leftrightarrow\ & \frac{\partial}{\partial z}[z^c - T(A(z))]\Big|_{z=\tilde{z}_0} = 0\ , \end{aligned}$$

meaning that $\tilde{z}_0$ is a zero of $z^c - T(A(z))$ with multiplicity larger than one, which is impossible according to lemma 1. □

Summarizing the theorems up to now, $W_1(z)$ has one dominant singularity, being the pole z_0. This dominant pole has multiplicity one and is equal to $A(\tilde{z}_0)$, with $\tilde{z}_0$ the only zero in $]1, R[$ of $z^c - T(A(z))$. As $\tilde{z}_0 \in \mathbb{R}$, it can be easily determined numerically, for instance with the bisection or the Newton-Raphson method [20].
Taking these findings into account, the first part of the packet delay (i.e., W_1) exhibits a geometric tail behavior and we obtain, similarly as in [3] and [4], the following dominant-pole approximation for the tail probabilities:

$$\begin{aligned} & \Pr[W_1 > w] \\ \approx\ & \frac{z_0^{-w-1}}{1-z_0}\frac{T(z_0)-1}{\mathrm{E}[A]}\frac{A\left(T(z_0)^{1/c}\right)-1}{\left(T(z_0)^{1/c}-1\right)^2}T(z_0)^{\frac{1}{c}-1} \\ & \times\frac{(z_0-1)\sum_{m=0}^{l-1}q_0(m)T(z_0)^{m/c} + \sum_{m=l}^{c-1}e(m)\left(T(z_0)-T(z_0)^{m/c}\right)}{c - A'\left(T(z_0)^{1/c}\right)T(z_0)^{\frac{1}{c}-1}T'(z_0)}\ . \end{aligned} \quad (8)$$

3.2. The calculation of $\Pr[W_2 > w]$

In order to calculate $\Pr[W_2 > w]$, we start from the following relation: (see also Fig. 1 for a brushup of the notations)

$$\Pr[W_2 > w] = \Pr\left[\left([Q_J + B] \bmod c\right) + 1 + X + \sum_{i=1}^{w} A_{J+i} < l\right] \quad , \qquad (9)$$

with "mod" the modulo operator. Indeed, the second part of the delay of a randomly tagged packet is larger than w if the sum of (a) the number of previously arrived packets that are served in the same batch as the tagged packet ($[Q_J + B] \bmod c$), (b) the tagged packet, (c) the number of packet arrivals during slot J after the tagged packet (X) and (d) the number of packet arrivals during the sequence of w slots following slot J ($\sum_{i=1}^{w} A_{J+i}$), is smaller than the threshold l. We transform this expression by means of the probability generating property of PGF's.
Since X and B are correlated, but independent of the other discrete random variables that appear in (9), we first compute $\mathrm{E}\left[x^{([Q_J+B] \bmod c)}\ x^X\right]$. We find, along the same lines as in our paper [11] p.6-8, that

$$\begin{aligned}
&\mathrm{E}\left[x^{([Q_J+B] \bmod c)}\ x^X\right] \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty} \Pr[Q_J + B = nc + m, X = k]\, x^m x^k \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\sum_{i=0}^{c-1} \Pr[Q_J + B = nc + m, X = k]\, x^i x^k \delta\langle m = i\rangle \quad , \quad (10)
\end{aligned}$$

with $\delta\langle . \rangle$ the Kronecker delta function, i.e.

$$\delta\langle m = i\rangle = \begin{cases} 1 & \text{if } m = i \ , \\ 0 & \text{if } m \neq i \ . \end{cases}$$

On account of the standard property

$$\delta\langle m = i\rangle = \frac{1}{c}\sum_{j=0}^{c-1} \varepsilon_j^{nc+m-i} \quad , \quad \forall n \in \mathbb{N} \ ,$$

(10) can be transformed into:

$$\begin{aligned}
&\mathrm{E}\left[x^{([Q_J+B] \bmod c)}\ x^X\right] \\
&= \frac{1}{c}\sum_{j=0}^{c-1}\sum_{i=0}^{c-1}\left(\frac{x}{\varepsilon_j}\right)^i \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\Pr\left[Q_J+B=nc+m, X=k\right]\varepsilon_j^{nc+m}x^k \ .
\end{aligned}$$

Owing to the IID character of the arrival process and the standard renewal argument (see e.g. [3]), we finally obtain

$$\mathrm{E}\left[x^{([Q_J+B] \bmod c)}\ x^X\right] = \frac{x^c-1}{c(x-1)}\sum_{j=0}^{c-1} Q(\varepsilon_j)\frac{A(\varepsilon_j)-A(x)}{\mathrm{E}\left[A\right](\varepsilon_j-x)}\frac{\varepsilon_j(x-1)}{x-\varepsilon_j} \ ,$$

with $Q(z)$ the PGF of the queue content at a random slot mark. In [12] p.8, formula (8), we have established the following expression for $Q(z)$:

$$\begin{aligned}
Q(z) &= \frac{1}{z^c - T(A(z))} \\
&\quad \times \left\{(z^c-1)\sum_{n=0}^{l-1} q_0(n)z^n + \frac{T(A(z))-1}{A(z)-1}\sum_{n=l}^{c-1} e(n)(z^c-z^n)\right\} \ ,
\end{aligned}$$

implying that $Q(\varepsilon_0)=1$ (since $\varepsilon_0=1$) and

$$Q(\varepsilon_j) = \frac{\sum_{n=l}^{c-1} e(n)(\varepsilon_j^n-1)}{A(\varepsilon_j)-1} \ , \quad 1 \le j \le c-1 \ .$$

Relying on this result, we find:

$$\mathrm{E}\left[x^{([Q_J+B] \bmod c)}\ x^X\right] = \frac{x^c-1}{c(x-1)} f(x) \ , \tag{11}$$

with

$$f(x) = \frac{1-A(x)}{\mathrm{E}\left[A\right](1-x)} + \sum_{j=1}^{c-1}\frac{A(\varepsilon_j)-A(x)}{\mathrm{E}\left[A\right](\varepsilon_j-x)}\frac{\varepsilon_j(x-1)}{x-\varepsilon_j}\frac{\sum_{n=l}^{c-1} e(n)(\varepsilon_j^n-1)}{A(\varepsilon_j)-1} \ ,$$

with the first term the one for $j=0$. The combination of (9), (11) and the probability generating property of PGF's produces

$$\begin{aligned}
\Pr\left[W_2 > w\right] &= \sum_{m=0}^{l-1}\frac{1}{m!}\frac{\partial^m}{\partial x^m}\mathrm{E}\left[x^{([Q_J+B] \bmod c)+1+X+\sum_{i=1}^{w} A_{J+i}}\right]\Big|_{x=0} \\
&= \sum_{m=1}^{l-1}\frac{1}{m!}\frac{\partial^m}{\partial x^m} x A(x)^w \frac{x^c-1}{c(x-1)} f(x)\Big|_{x=0} \ .
\end{aligned}$$

After some mathematical manipulations, this can be transformed into

$$\begin{aligned}\Pr\left[W_2 > w\right] &= \sum_{m=0}^{l-2} \frac{1}{m!} \frac{\partial^m}{\partial x^m} A(x)^w \frac{x^c - 1}{c(x-1)} f(x) \Big|_{x=0} \\ &= \frac{1}{c} \sum_{m=0}^{l-2} \sum_{k=0}^{m} \frac{1}{k!(m-k)!} \frac{\partial^k}{\partial x^k} \frac{x^c - 1}{x - 1} \Big|_{x=0} \frac{\partial^{m-k}}{\partial x^{m-k}} A(x)^w f(x) \Big|_{x=0} \quad .\end{aligned}$$

Invoking

$$\frac{x^c - 1}{x - 1} = \sum_{n=0}^{c-1} x^n \quad ,$$

yields

$$\frac{\partial^k}{\partial x^k} \frac{x^c - 1}{x - 1} \Big|_{x=0} = k! \quad ,$$

for all $k < c$. Hence

$$\begin{aligned}\Pr\left[W_2 > w\right] &= \frac{1}{c} \sum_{m=0}^{l-2} \sum_{k=0}^{m} \frac{1}{(m-k)!} \frac{\partial^{m-k}}{\partial x^{m-k}} A(x)^w f(x) \Big|_{x=0} \\ &= \frac{1}{c} \sum_{k=0}^{l-2} \frac{l-1-k}{k!} \frac{\partial^k}{\partial x^k} A(x)^w f(x) \Big|_{x=0} \quad . \qquad (12)\end{aligned}$$

Remark 1. *When $l = c$, the sum over j in $f(x)$ vanishes, so that we find that $\Pr\left[W_2 > w\right]$ is not influenced by the distribution of the service lengths in that case, even not by the mean value.*

Formula (12) can be implemented in a mathematical program such as matlab. This procedure suffers from the drawback that high-order derivatives may have to be computed, which causes a considerable reduction in speed and even is infeasible if l and c are quite large. Therefore, we now deduce an approximation for $\Pr\left[W_2 > w\right]$, whereby no derivatives have to be taken.

The PGF associated with W_2, $W_2(z)$, is extracted from (12) by multiplying both sides of the equation by z^w and taking the sum over all values of w:

$$W_2(z) = 1 + \frac{z-1}{c} \sum_{m=0}^{l-2} \frac{l-1-m}{m!} \frac{\partial^m}{\partial x^m} \frac{f(x)}{1 - zA(x)} \Big|_{x=0} \quad . \qquad (13)$$

From this equation, it is clear that $z = 1/A(0)$ is the dominant pole of $W_2(z)$ and that it has multiplicity $l-1$. We now determine the behavior of $W_2(z)$ about the dominant singularity. The m-th ($m \geq 0$) derivative of $f(x)/(1-zA(x))$ can be written as

$$\frac{\partial^m}{\partial x^m}\frac{f(x)}{1-zA(x)} = \sum_{j=0}^{m}\frac{C_{m,j}(z,x)}{[1-zA(x)]^{j+1}} \ , \tag{14}$$

whereby $C_{m,j}(z,x)$ are functions of z and x that have no factor $1-zA(x)$. As opposed to $C_{m,j}(z,x)$ for $j \neq m$, $C_{m,m}(z,x)$ is relatively easy to calculate:

$$C_{m,m}(z,x) = m!f(x)z^m A^{'}(x)^m \ .$$

The substitution of (14) in (13) yields

$$W_2(z) = 1 + \frac{z-1}{c}\sum_{m=0}^{l-2}\frac{l-1-m}{m!}\sum_{j=0}^{m}\frac{C_{m,j}(z,0)}{[1-zA(0)]^{j+1}} \ . \tag{15}$$

Consequently, if we retain the simple most dominant term from this expression, we find that $W_2(z)$ is proportional to

$$W_2(z) \sim \frac{z-1}{c}f(0)\frac{(zA^{'}(0))^{l-2}}{[1-zA(0)]^{l-1}} \ ,$$

in a neighborhood of $z = 1/A(0)$. The dominant-pole approximation thus yields:

$$\Pr\left[W_2 > w\right] \approx w^{l-2}A(0)^w\frac{f(0)}{c(l-2)!}\left(\frac{A^{'}(0)}{A(0)}\right)^{l-2} \ .$$

However, we notice that this increases as w increases, for $0 \leq w \leq (2-l)/\ln(A(0))$. When, for instance $l = 10$ and $A(0) = e^{-0.5}$, $(2-l)/\ln(A(0))$ equals 22, which indicates that the approximation is probably inaccurate for w between 0 and 22 (and even for larger w-values) as $\Pr\left[W_2 > w\right]$ is obviously a monotonically decreasing function. We therefore propose a more accurate approximation formula. Mark that we only retained the term with $j = m = l-2$ about $z = 1/A(0)$ in (15), as it produces the largest power of $1-zA(0)$ in the denominator. Instead of only retaining this term, we take all the terms into account for which $j = m$. We thus retain for every m the

term that produces the largest power of $1 - zA(0)$ in the denominator. We thus take advantage of the fact that we can easily calculate $C_{m,m}(z,x)$ for all m. Hence, $W_2(z)$ transforms into

$$W_2(z) \sim \frac{z-1}{c} \sum_{m=0}^{l-2} \frac{(l-1-m) f(0) z^m A^{'}(0)^m}{[1-zA(0)]^{m+1}} \quad . \tag{16}$$

Next, $1/[1-zA(0)]^{m+1}$ can be rewritten as follows:

$$\begin{aligned} \frac{1}{[1-zA(0)]^{m+1}} &= \frac{1}{m! A(0)^m} \frac{\mathrm{d}^m}{\mathrm{d}z^m} \frac{1}{1-A(0)z} \\ &= \frac{1}{m! A(0)^m} \frac{\mathrm{d}^m}{\mathrm{d}z^m} \sum_{w=0}^{\infty} [A(0)z]^w \\ &= \frac{1}{m! A(0)^m} \sum_{w=m}^{\infty} A(0)^w z^{w-m} \frac{w!}{(w-m)!} \quad . \end{aligned} \tag{17}$$

The second step requires that $|A(0)z| < 1$, which is satisfied for z approaching $1/A(0)$ from the left. The substitution of (17) in (16) produces:

$$\begin{aligned} \frac{W_2(z)-1}{z-1} &\sim \frac{f(0)}{c} \sum_{m=0}^{l-2} A^{'}(0)^m (l-1-m) \sum_{w=m}^{\infty} z^w \frac{w!}{m!(w-m)!} A(0)^{w-m} \\ &= \frac{f(0)}{c} \sum_{w=0}^{\infty} z^w \sum_{m=0}^{\min(l-2,w)} A^{'}(0)^m (l-1-m) \binom{w}{m} A(0)^{w-m} \quad , \end{aligned}$$

so that the approximation formula reads:

$$\Pr[W_2 > w] \approx \frac{f(0)}{c} \sum_{m=0}^{\min(l-2,w)} A^{'}(0)^m (l-1-m) \binom{w}{m} A(0)^{w-m} \quad . \tag{18}$$

Note that for large w, formula (18) becomes a sum from 0 to $l-2$. We further point out that the binomial coefficient causes no difficulties, since efficient routines exist to calculate them, even for large w.

Remark 2. *As $z = 1/A(0)$ is a pole with multiplicity larger than 1 if $l \geq 3$, W_2 does not exhibit a purely geometric tail behaviour.*

Remark 3. *Note that this approach is not suited for cases whereby $A^{'}(0) = 0$, as only the term corresponding to $m = 0$ in (16) differs from 0. In these cases, additional terms with $j < m$ must be taken into account in (15).*

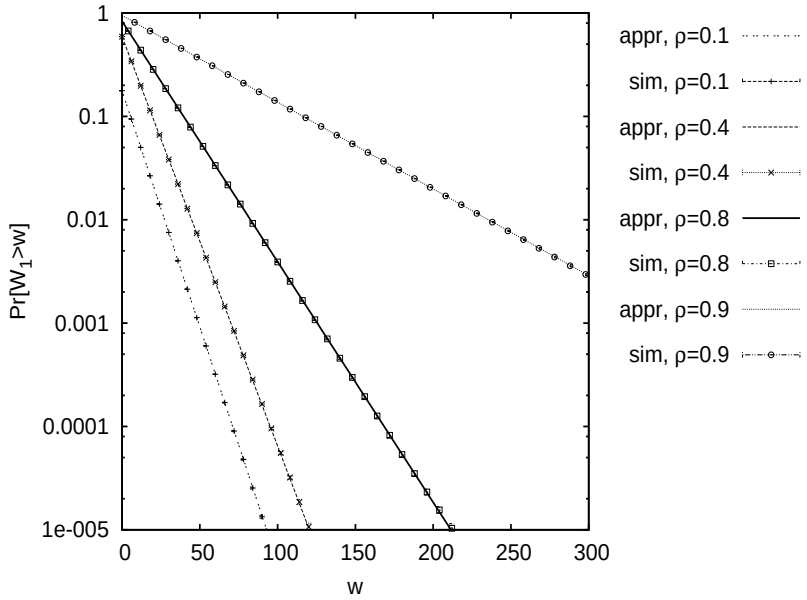
4. Accuracy of the formulas

In this section, we evaluate the accuracy of our approach. First, we study formula (8) for $\Pr[W_1 > w]$. Then, we focus on approximation (18) for $\Pr[W_2 > w]$ and finally the accuracy of the bounds for $\Pr[W > w]$ is covered.

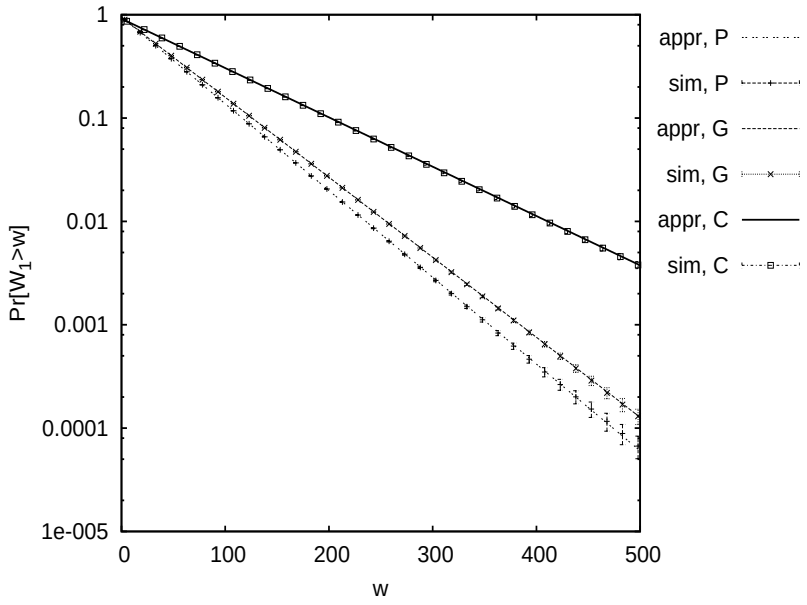
In Figures 2-4, approximation (8) as well as simulated values[2] for $\Pr[W_1 > w]$ are depicted versus w for various combinations of server capacities c, service thresholds l, loads ρ and several distributions for the number of packet arrivals (Poisson $A(z) = e^{\mathrm{E}[A](z-1)}$; Geometric $A(z) = 1/(1 + \mathrm{E}[A] - \mathrm{E}[A]\, z)$; C-center $A(z) = 1 - \mathrm{E}[A]/c + \mathrm{E}[A]/(2c)(z^{c-1} + z^{c+1})$) and service times (Geometric $T(z) = z/[\mathrm{E}[T] + (1 - \mathrm{E}[T])z]$; 25 $T(z) = 1 - \mathrm{E}[T]/25 + \mathrm{E}[T]/25 z^{25}$ with $\mathrm{E}[T] = 5$ or 10). We observe that approximation formula (8) is accurate, even for relatively small values of w.
The figures further exhibit that higher loads lead to larger tail probabilities of the first part of the packet delay. In addition, the distributions of the service times and the amount of per-slot arrivals have an undeniable impact. Next, we perceive that although the load remains equal (and thus the mean arrival rate $\mathrm{E}[A]$ increases) a larger server capacity c has a positive influence on the tail probabilities. Finally, the service threshold has a negligible impact on $\Pr[W_1 > w]$.

Next, approximation (18) and exact formula (12) for $\Pr[W_2 > w]$ are depicted versus w in Figures 5-7 for various settings of the system parameters. The figures highlight that there might be a discernable error for small values of w, whereas it becomes small for larger values of w. We also observe that although the relative error might sometimes be quite large, the order of magnitude of $\Pr[W_2 > w]$ is well approximated, which is sufficient for e.g. the purpose of assessing the QoS of a voice conversation, even for smaller values of w. In a voice conversation, one typically experiences an acceptable quality when the probability that the delay of a voice packet does not exceed 150 ms is smaller than 10^{-2} (see e.g. [15]). The approximation formula is also a lot faster than exact formula (12). In order to give an idea: the computing time to calculate $\Pr[W_2 > 100]$ in Fig. 7 (e) when $l = 10$, equals 6.33s via

[2]We have depicted the confidence intervals resulting from 20 simulations whereby each simulation generates the delay for 10^9 packets

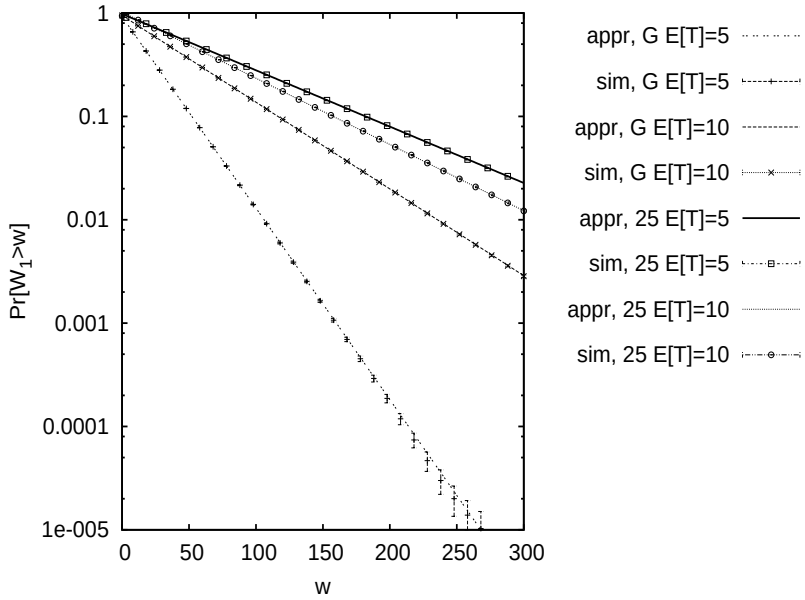


(a) several loads; $c = 10$, $l = 5$, Poisson arrivals, geometric services mean 10

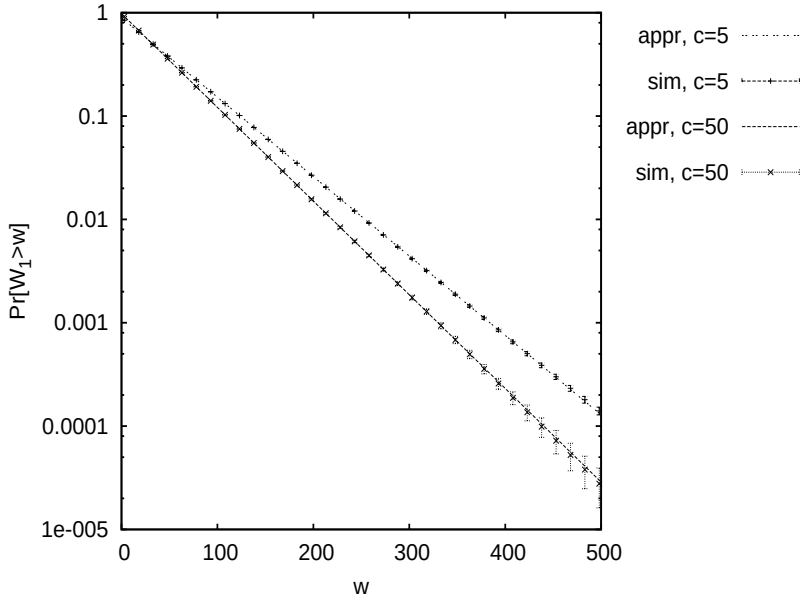


(b) several $A(z)$'s; $c = 10$, $l = 5$, $\rho = 0.9$, geometric services mean 10

Figure 2: Evaluation of approximation formula (8) for $\Pr[W_1 > w]$ (1)

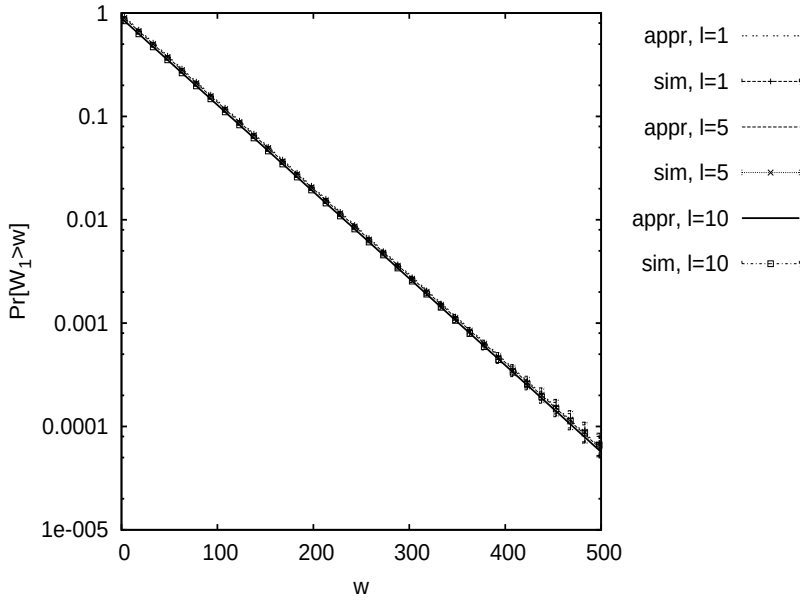


(c) several $T(z)$'s; $c = 10$, $l = 5$, $\rho = 0.9$, Poisson arrivals



(d) several c's; $l = 5$, $\rho = 0.9$, Poisson arrivals, geometric services mean 10

Figure 3: Evaluation of approximation formula (8) for $\Pr[W_1 > w]$ (2)

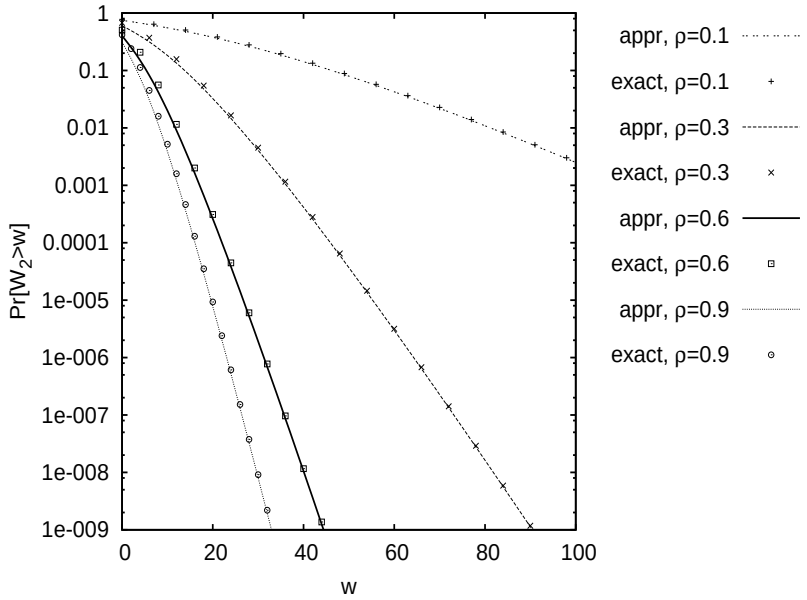


(e) several l's; $c = 10$, $\rho = 0.9$, Poisson arrivals, geometric services mean 10

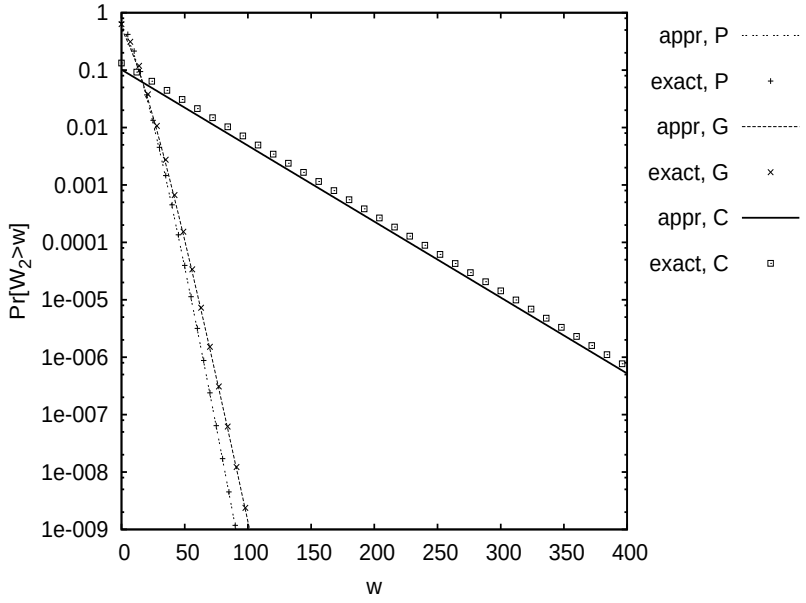
Figure 4: Evaluation of approximation formula (8) for $\Pr[W_1 > w]$ (3)

formula (12), whereas it only takes 0.25s via approximation formula (18). Ultimately, for increasing values of c and l, the calculation of (12) becomes unfeasible, which highlights the necessity of the approximation formula.
Further, the figures highlight that $A(z)$, $T(z)$, $\mathrm{E}[A]$, l and c have a significant impact on the second part of the packet delay, whereas the opposite is true for $\mathrm{E}[T]$. However, $\mathrm{E}[T]$ has a slight influence on $\Pr[W_2 > w]$ through the unknowns $e(n)$. Only when $l = c$, $\mathrm{E}[T]$ (and even $T(z)$) has no influence at all (see remark 1).

Let us now investigate the accuracy of bounds (3) and (4) for $\Pr[W > w]$. We have therefore plotted these bounds versus w in Figures 8-10 for a broad range of system parameters. We perceive that the bounds nearly coincide, except for some values of the load ρ. In order to investigate this issue further, the bounds for $\Pr[W > w]$ are depicted versus the load for several examples in Fig. 11. We observe that $\Pr[W > w]$ is the largest when $\rho \to 0$ and $\rho \to 1$ and that the bounds nearly coincide in these cases. Indeed, when $\rho \to 0$, few packets arrive, leading to a very long

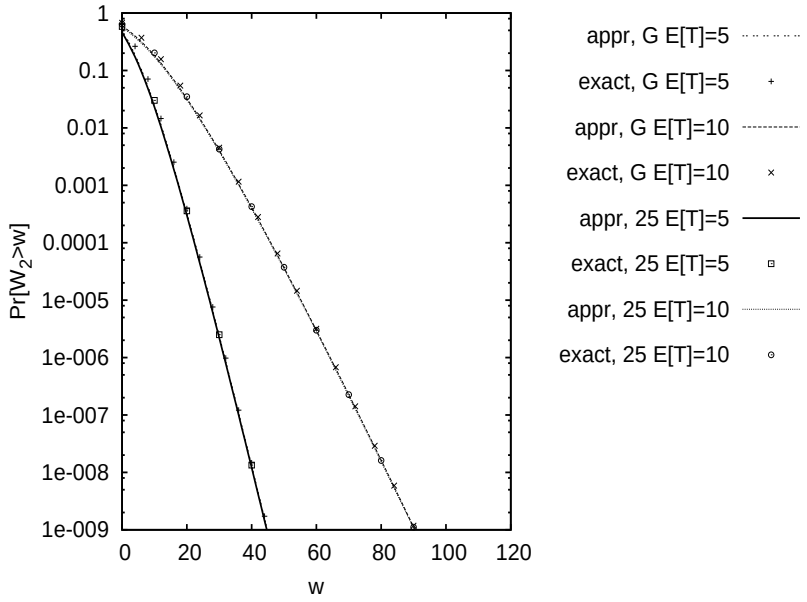


(a) several loads; $c = 10$, $l = 5$, Poisson arrivals, geometric services mean 10

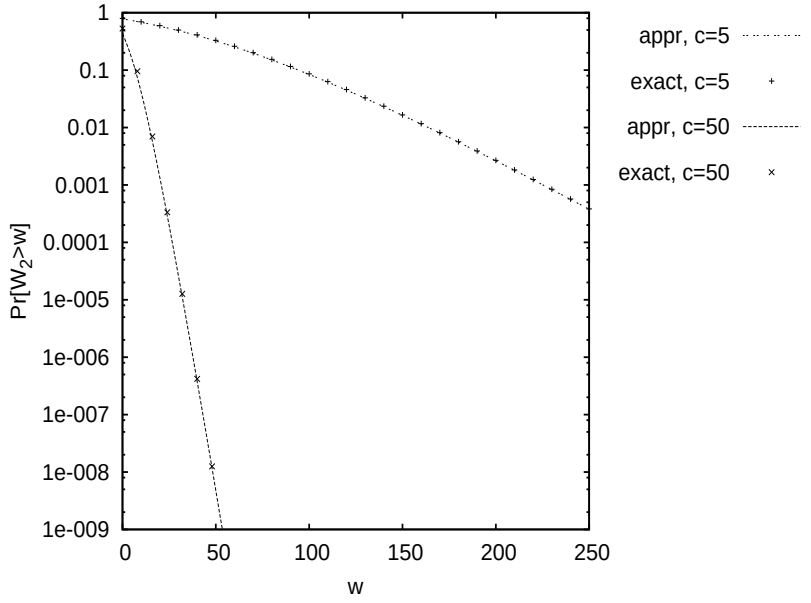


(b) several $A(z)$'s; $c = 10$, $l = 5$, $\rho = 0.3$, geometric services mean 10

Figure 5: Evaluation of approximation formula (18) for $\Pr[W_2 > w]$ (1)

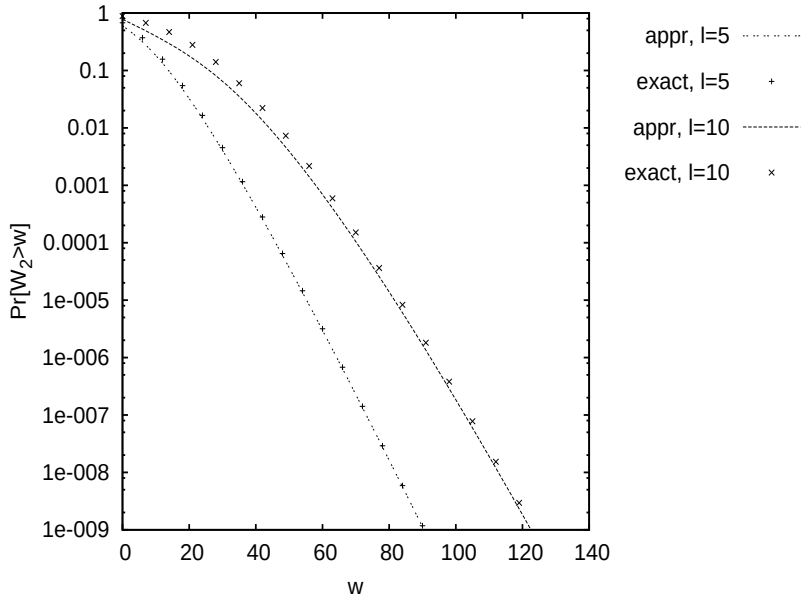


(c) several $T(z)$'s; $c = 10$, $l = 5$, $\rho = 0.3$, Poisson arrivals



(d) several c's; $l = 5$, $\rho = 0.3$, Poisson arrivals, geometric services mean 10

Figure 6: Evaluation of approximation formula (18) for $\Pr[W_2 > w]$ (2)



(e) several l's; $c = 10$, $\rho = 0.3$, Poisson arrivals, geometric services mean 10

Figure 7: Evaluation of approximation formula (18) for $\Pr[W_2 > w]$ (3)

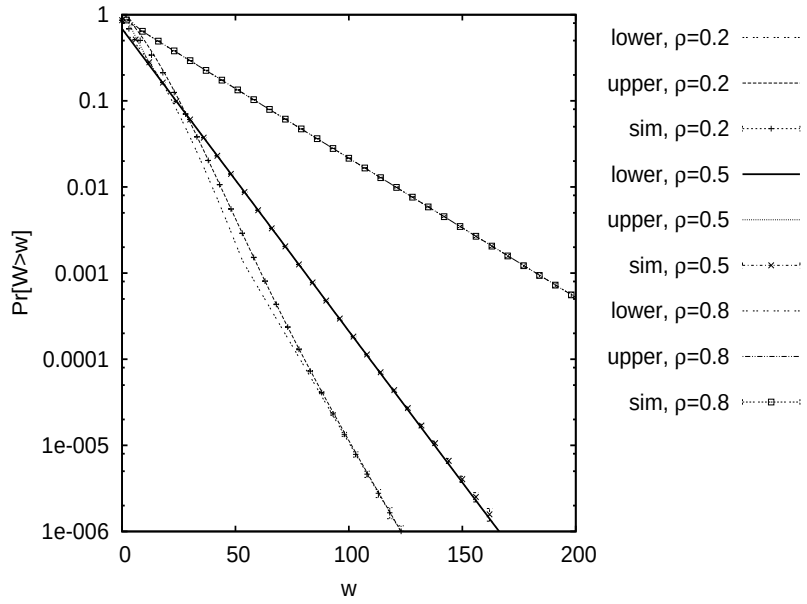
second part and a negligible short first part of the delay, whereas when $\rho \to 1$, the opposite holds. We also observe that $\Pr[W > w]$ decreases until its minimum, whereafter it increases again. In addition, the largest difference between the bounds appears in the neighbourhood of the minimum of the curves. This can be explained as follows: when ρ increases, $\Pr[W_1 > w]$ increases, whereas $\Pr[W_2 > w]$ decreases. Consequently, the difference between the bounds, $\min(\Pr[W_1 > w], \Pr[W_2 > w])$, is the largest when $\Pr[W_1 > w] = \Pr[W_2 > w]$. In that case, we learn from (3) and (4) that the upper bound is (roughly) twice as large as the lower bound.

Although we have demonstrated that the bounds nearly overlap, we have to bear in mind that these bounds rely on expressions for $\Pr[W_1 > w]$ and $\Pr[W_2 > w]$, for which we also use approximations. As these are very accurate for large w, we expect that this accumulation of errors remains small. In order to verify this, we have also depicted simulated values[3] of $\Pr[W > w]$

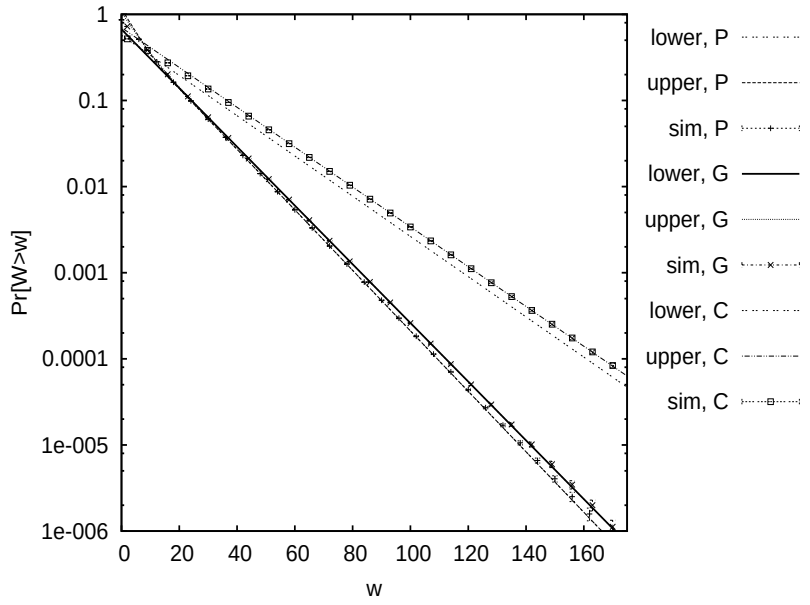
[3]We have depicted the confidence intervals resulting from 20 simulations whereby each

in figures 8-11. We observe that the curves are very close to the simulated values as anticipated. We can thus conclude that bounds (3) and (4) together with approximations formulas (8) and (18) for $\Pr[W_1 > w]$ and $\Pr[W_2 > w]$ are very accurate. These are for instance useful to assess the probability that a delay-sensitive packet (for instance a voice packet) experiences an excessive delay in an output buffer of a node in a UWB WPAN.

simulation generates the delay for 10^9 packets

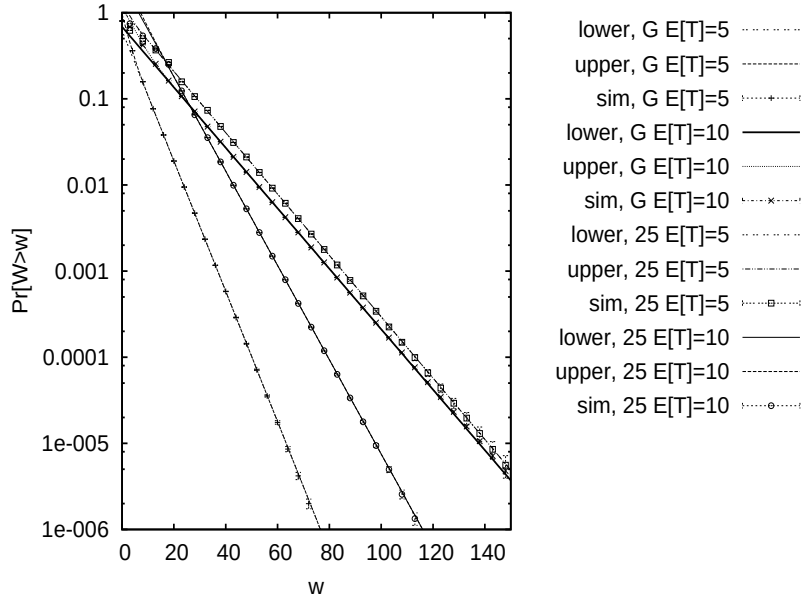


(a) several loads; $c = 10$, $l = 5$, Poisson arrivals, geometric services mean 10

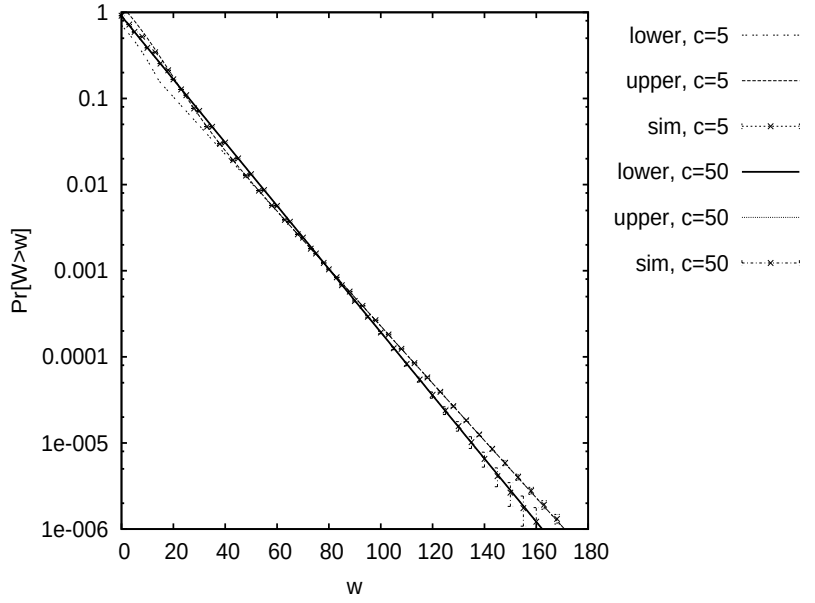


(b) several $A(z)$'s; $c = 10$, $l = 5$, $\rho = 0.5$, geometric services mean 10

Figure 8: Evaluation of bounds (3) and (4) for $\Pr[W > w]$ versus w (1); approximation formula (18) for $\Pr[W_2 > w]$ is used except in (b) for C-centered arrivals, where we have adopted formula (12) because $A^{'}(0) = 0$

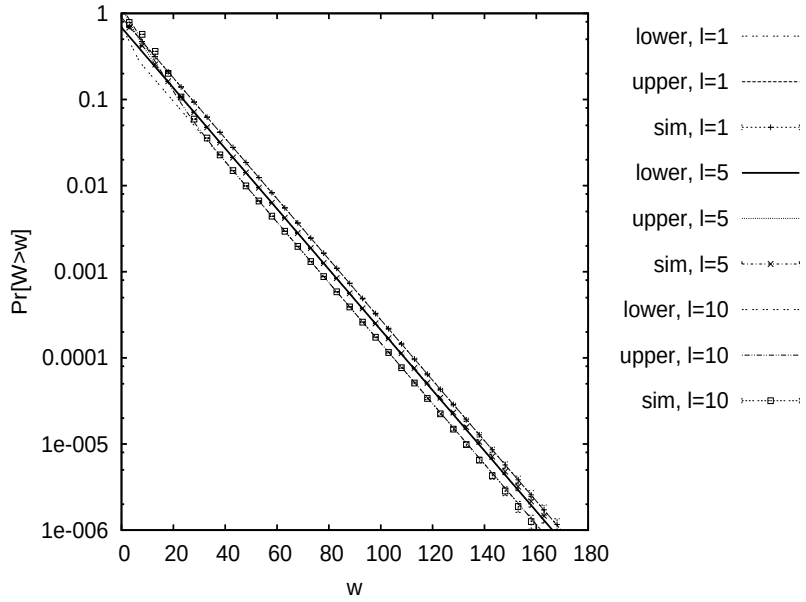


(c) several $T(z)$'s; $c = 10$, $l = 5$, $\rho = 0.5$, Poisson arrivals



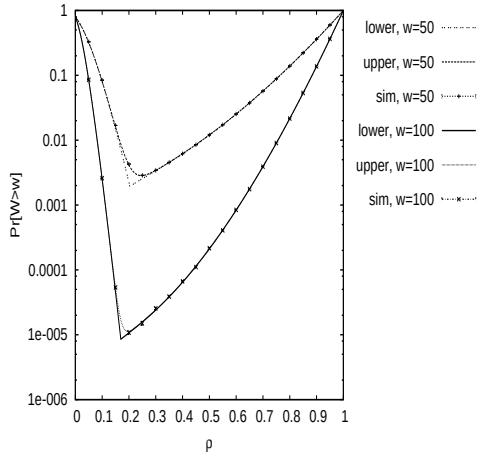
(d) several c's; $l = 5$, $\rho = 0.5$, Poisson arrivals, geometric services mean 10

Figure 9: Evaluation of bounds (3) and (4) for $\Pr[W > w]$ versus w (2); approximation formula (18) for $\Pr[W_2 > w]$ is used

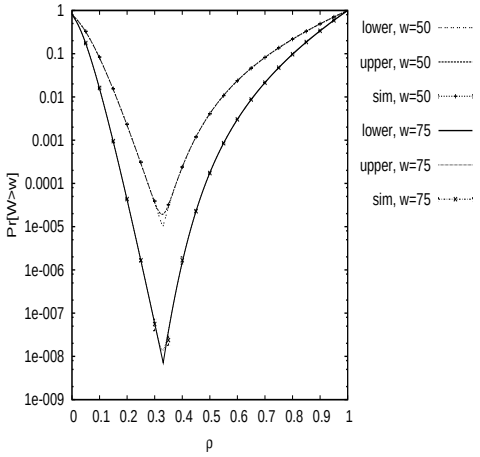


(e) several l's; $c = 10$, $\rho = 0.5$, Poisson arrivals, geometric services mean 10

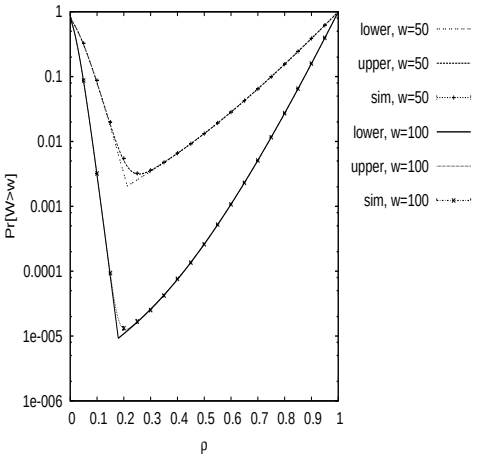
Figure 10: Evaluation of bounds (3) and (4) for $\Pr[W > w]$ versus w (3); approximation formula (18) for $\Pr[W_2 > w]$ is used



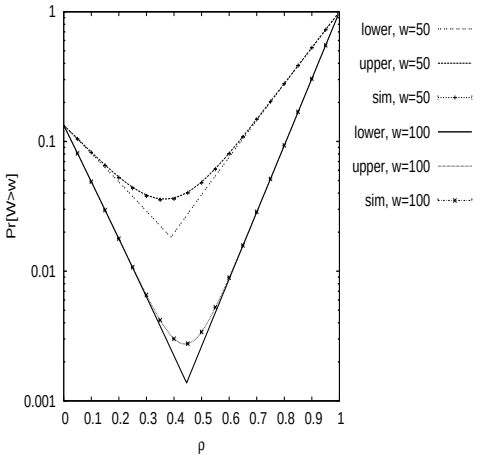
(a) Poisson arrivals, geometric services



(b) Poisson arrivals, 1 or 25 slots service



(c) geometric arrivals, geometric services



(d) c-centered arrivals, geometric services

Figure 11: Evaluation of bounds (3) and (4) for $\Pr[W > w]$ versus the load; approximation formula (18) for $\Pr[W_2 > w]$ is used, except in (d) for C-centered arrivals, where we have adopted formula (12) because $A^{'}(0) = 0$; $c = 10$, $l = 5$, $\mathrm{E}[T] = 10$

[1] Arumuganathan, R., Jeyakumar, S.: Steady state analysis of a bulk queue with multiple vacations, setup times with N-policy and closedown times. Appl. Math. Model. 29, 972–986 (2005)

[2] Bellalta, B.: A Queueing Model for the Non-continuous Frame Assembly Scheme in Finite Buffers. Proc. 16th Int. Conf. Anal. Stoch. Model. Tech. Appl. (ASMTA 2009), Madrid, June 9-12, pp.219–233 (2009)

[3] Bruneel, H., Kim, B.G.: Discrete-Time Models for Communication Systems Including ATM. Kluwer Academic Publishers, Boston/Dordrecht/London (1993)

[4] Bruneel, H., Steyaert, B., Desmet, E., Petit, G.H.: Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. Eur. J. Oper. Res. 76, 563–572 (1994)

[5] Chang, S.H., Choi, D.W.: Performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and vacations. Comp. Oper. Res. 32, 2213–2234 (2005)

[6] Chang, S.H., Takine, T.: Factorization and Stochastic Decomposition Properties in Bulk Queues with Generalized Vacations. Queueing Syst. 50, 165–183 (2005)

[7] Chaudhry, M.L., Templeton, J.G.C.: A first course in bulk queues. John Wiley & Sons (1983)

[8] Claeys, D., Walraevens, J., Laevens, K., Bruneel, H.: A Discrete-time Queueing Model with a Batch Server Operating under the Minimum Batch Size Rule. Proc. 7th Int. Conf. on Next Gener. Teletraffic and Wired/Wirel. Adv. Netw. (NEW2AN 2007), St. Petersburg, September 10-14, pp.248–259 (2007)

[9] Claeys, D., Laevens, K., Walraevens, J., Bruneel, H.: Delay in a discrete-time queueing model with batch arrivals and batch services. Proc. Fifth Int. Conf. Inf. Tech.: New Gener. (ITNG 2008), Las Vegas, Nevada, April 7-9, pp.1040–1045 (2008)

[10] Claeys, D., Walraevens, J., Laevens, K., Bruneel, H.: Delay analysis of two batch-service queueing models with batch arrivals: $Geo^X/Geo^c/1$. 4OR 8(3), 255–269 (2010)

[11] Claeys, D., Laevens, K., Walraevens, J., Bruneel, H.: Complete characterisation of the customer delay in a queueing system with batch arrivals and batch service. Math. Meth. Oper. Res. 72(1), 1–23 (2010)

[12] Claeys, D., Walraevens, J., Laevens, K., Bruneel, H.: Analysis of threshold-based batch-service queueing systems with batch arrivals and general service times. accepted in Perform. Eval.

[13] Cohen, J.W.: The single server queue. North-Holland, Amsterdam; Wiley Interscience, New York (1969)

[14] Dagsvik, J.: The general bulk queue as a matrix factorisation problem of the Wiener-Hopf type. Part 1. Adv. Appl. Prob. 7(3), 636–646 (1975)

[15] De Vleeschauwer, D., Van Moffaert, A., Büchli, M., Petit, G.H., Steyaert, B., Bruneel, H.: Determining the tolerable load generated by a set of packet-based phones on a multiplexing node. Proc. 17th Int. Teletraffic Congr. (ITC 17), Salvador da Bahia, Brazil, December 2-7 (2001)

[16] Downton, F.: Waiting Time in Bulk Service Queues. J. R. Stat. Soc., Series B (Methodological) 17(2), 256–261 (1955)

[17] Dümmler, M.A., Schömig, A.K.: Using discrete-time analysis in the performance evaluation of manufacturing systems. Proc. 1999 Int. Conf. Semicond. Manuf. Oper. Model. Simul. (SMOMS '99), San Francisco, California, January 18-20 (1999)

[18] Goswami, V., Mohanty, J.R., Samanta, S.K.: Discrete-time bulk-service queues with accessible and non-accessible batches. Appl. Math. Comput. 182, 898–906 (2006)

[19] Gupta, U.C., Goswami, V.: Performance analysis of finite buffer discrete-time queue with bulk service. Comp. Oper. Res. 29, 1331–1341 (2002)

[20] Hildebrand, F.B.: Introduction to numerical analysis (2nd edition). McGraw-Hill (1974)

[21] Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: Analytic Computation Schemes for the Discrete-Time Bulk Service Queue. Queueing Syst. 50, 141-163 (2005)

[22] Keilson, J.: The general bulk queue as a Hilbert problem. J. R. Stat. Soc., Series B (Methodological) 24(2), 344–358 (1962)

[23] Kim, N.K., Chaudhry, M.L.: Equivalences of Batch-Service Queues and Multi-Server Queues and Their Complete Simple Solutions in Terms of Roots. Stoch. Anal. Appl. 24, 753–766 (2006)

[24] Lee, H.W., Lee, S.S., Chae, K.C.: A Fixed-size Batch Service Queue with Vacations. J. Appl. Math. Stoch. Anal. 9, 205–219 (1996)

[25] Lu, K., Wu, D., Fang, Y., Qiu, R.C.: Performance Analysis of A Burst-Frame-Based MAC Protocol for Ultra-Wideband Ad Hoc Networks. Proc. IEEE Int. Conf. Comm. 2005 (ICC 2005), Seoul, May 16-20, Vol.5, pp.2937–2941 (2005)

[26] Medhi, J.: Waiting Time Distributions in a Poisson Queue with a General Bulk Service Rule. Manag. Sci. 21(2), 777–782 (1975)

[27] Miller, R.G.: A contribution to the theory of bulk queues. J. R. Stat. Soc., Series B (Methodological) 21(2), 320–337 (1959)

[28] Powell, W.B., Humblet, P.: The bulk service queue with a general control strategy: theoretical analysis and a new computational procedure. Oper. Res. 34(2), 267–275 (1986)

[29] Samanta, S.K., Chaudhry, M.L., Gupta, U.C.: Discrete-time $Geo^X|G^{(a,b)}|1|N$ queues with single and multiple vacations. Math. Comp. Model. 45, 93–108 (2007)

[30] Sikdar, K., Gupta, U.C.: Analytic and numerical aspects of batch service queues with single vacation. Comp. Oper. Res. 32, 943–966 (2005)

[31] Steyaert, B.: Analysis of generic discrete-time buffer models with irregular packet arrival patterns. http://biblio.ugent.be/record/471285, Phd-thesis, Ghent University (promoter: Herwig Bruneel) (2008)

[32] Yi, X.W., Kim, N.K., Yoon, B.K., Chae, K.C.: Analysis of the queue-length distribution for the discrete-time batch-service $Geo^X|G^{a,Y}|1|K$ queue. Eur. J. Oper. Res. 181, 787–792 (2007)

[33] Zhao, Y.Q., Campbell, L.L.: Equilibrium probability calculations for a discrete-time bulk queue model. Queueing Syst. 22, 189–198 (1996)